COVID-lab1-ch6.R

karanYsingh

2021-03-16

```r
# Loading data
knitr::opts_chunk$set(cache =TRUE)
covid_19_data = read.csv("C:\\Users\\91828\\Documents\\Rlab\\Covid\\Cleaned_Data.csv", header = TRUE)

# covid_19_data = covid_19_data[0:10000,]

library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.0.4
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-1
```

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.3
```

```
## Warning: package 'tibble' was built under R version 4.0.3
```

```
## Warning: package 'tidyr' was built under R version 4.0.3
```

```
## Warning: package 'readr' was built under R version 4.0.3
```

```
## Warning: package 'purrr' was built under R version 4.0.3
```

```
## Warning: package 'dplyr' was built under R version 4.0.3
```

```
## Warning: package 'stringr' was built under R version 4.0.3
## Warning: package ’forcats’ was built under R version 4.0.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tidyr::pack()   masks Matrix::pack()
## x tidyr::unpack() masks Matrix::unpack()
```

```r
library(caret)
```

```
## Warning: package ’caret’ was built under R version 4.0.3
```

```
## Loading required package: lattice
```

```
##
## Attaching package: ’caret’
```

```
## The following object is masked from ’package:purrr’:
##
##     lift
```

```r
# Load the data and remove NAs
covid_19_data <- na.omit(covid_19_data)

# Inspect the data
sample_n(covid_19_data, 3)
```

```
##   Fever Tiredness Dry.Cough Difficulty.in.Breathing Sore.Throat None_Sympton
## 1     0         0         0                       1           1            0
## 2     0         0         0                       1           0            0
## 3     0         0         1                       0           0            0
##   Pains Nasal.Congestion Runny.Nose Diarrhea None_Experiencing Age_0.9
## 1     0                0          1        1                 0       0
## 2     0                1          1        0                 0       0
## 3     0                1          0        0                 0       1
##   Age_10.19 Age_20.24 Age_25.59 Age_60. Gender_Female Gender_Male
## 1         0         0         1       0             0           1
## 2         0         0         0       1             0           0
## 3         0         0         0       0             1           0
##   Gender_Transgender Severity_Mild Severity_Moderate Severity_None
## 1                  0             1                 0             0
## 2                  1             0                 1             0
## 3                  0             0                 0             1
##   Severity_Severe Contact_Dont.Know Contact_No Contact_Yes   Country
## 1               0                 1          0           0 Other-EUR
## 2               0                 0          1           0      Iran
## 3               0                 0          1           0     Italy
```

```r
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.0.4
```

```r
#nvmax = 27 to include all 27 parameters
regfit.full = regsubsets(Difficulty.in.Breathing~.,covid_19_data,nvmax = 27)
```

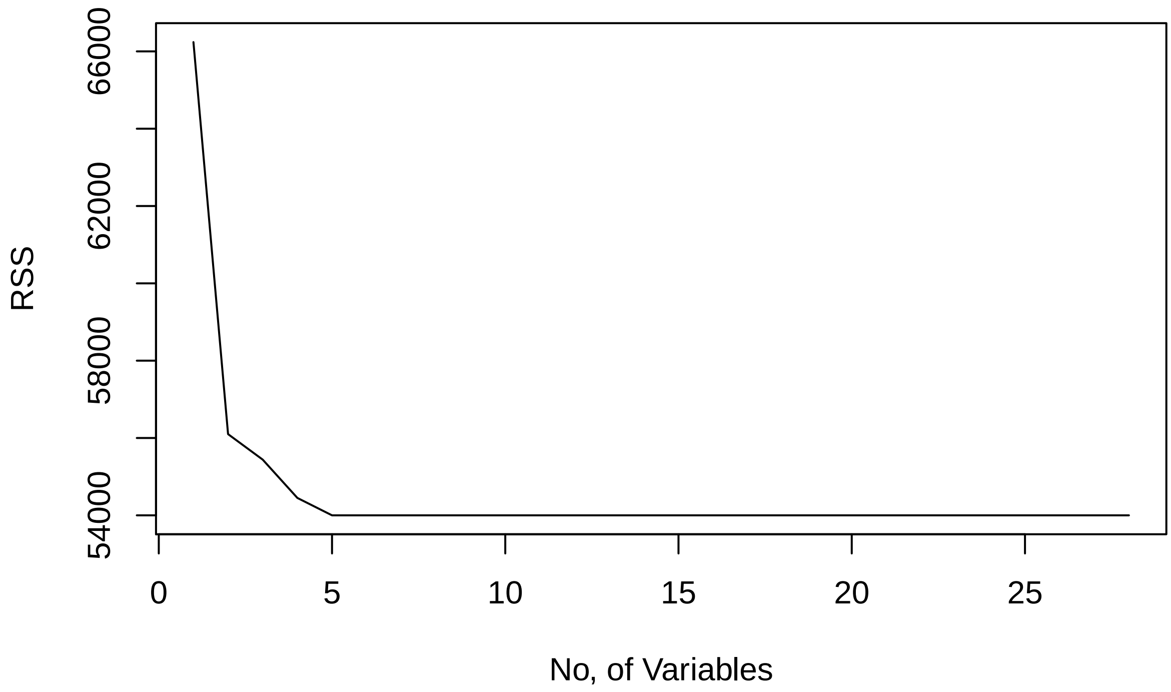
```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 4 linear dependencies found
```

```
## Reordering variables and trying again:
```

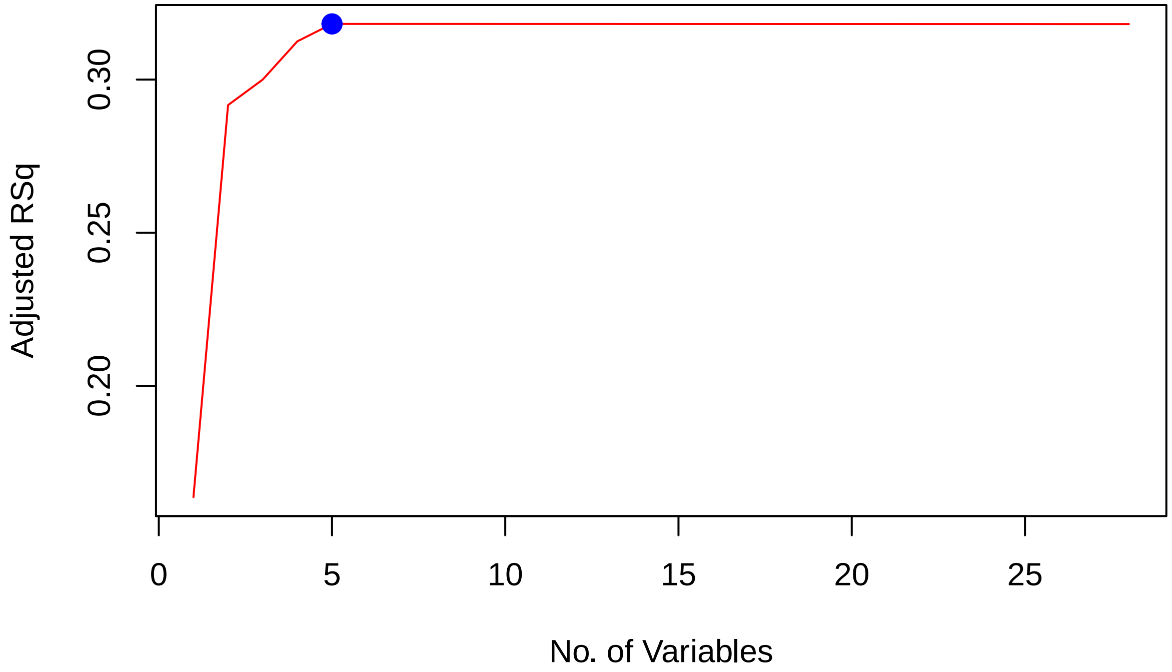
```r
covid.summary = summary(regfit.full)
names(covid.summary)
```

```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

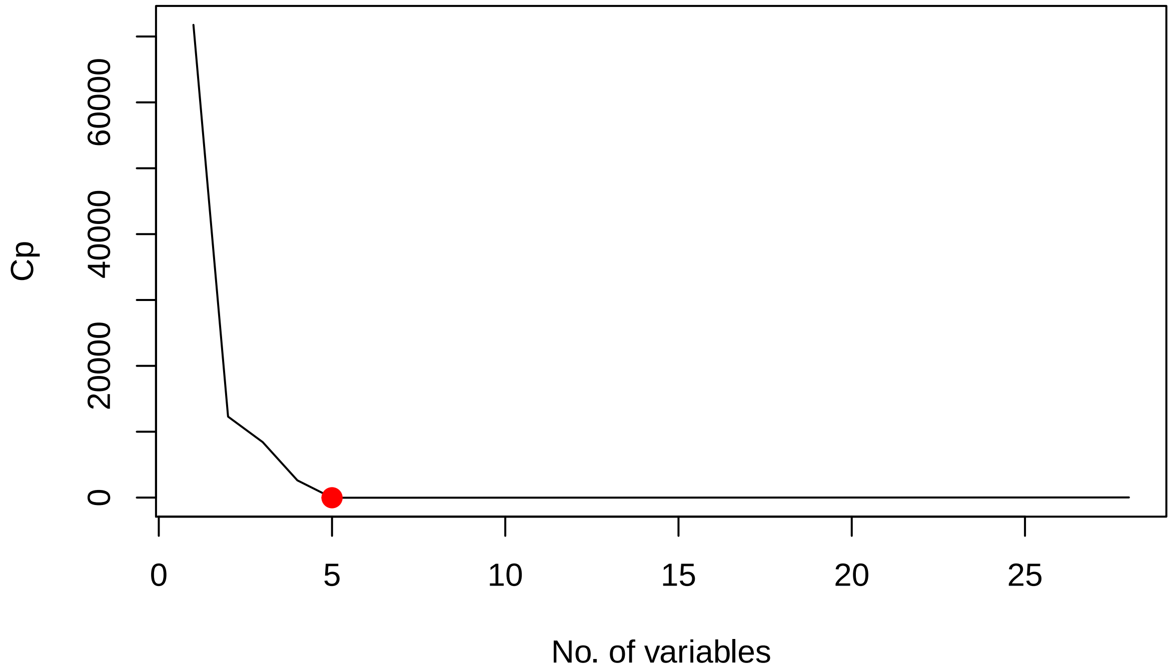
```r
# par(mfrow=c(2,2))
plot(covid.summary$rss,xlab="No, of Variables",ylab="RSS",type="l")
```



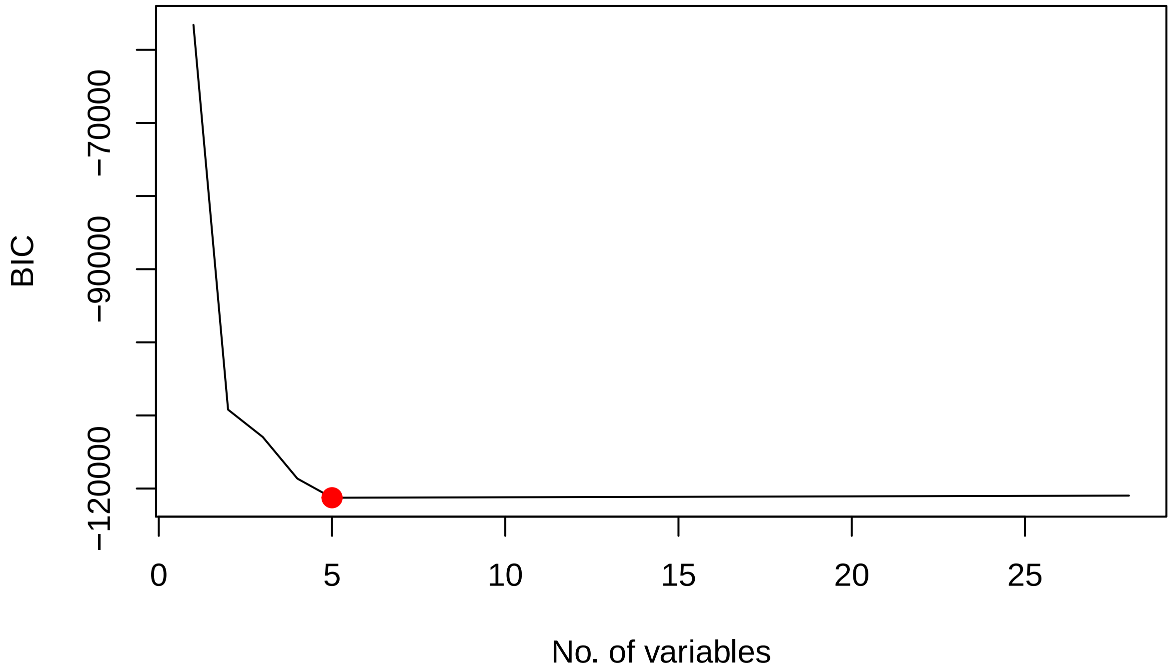
```
plot(covid.summary$adjr2,xlab="No. of Variables",ylab="Adjusted RSq",type="l",col="red")

#RSS always generally decreases(and Rsquared increases) as more predictors are used,
#From the plot It can’t be inferred that all predictors are significant.
#Blue point indicates the model with max RSquared.
maxRsq = which.max(covid.summary$adjr2)
points(maxRsq,covid.summary$adjr2[maxRsq],col="blue",cex=2,pch=20)
```



```
#Blue Dots indicate smallest statistics in Cp and BIC plots.
plot(covid.summary$cp,xlab="No. of variables",ylab="Cp",type="l")
minCp = which.min(covid.summary$cp)
points(minCp,covid.summary$cp[minCp],col="red",cex=2,pch=20)
```



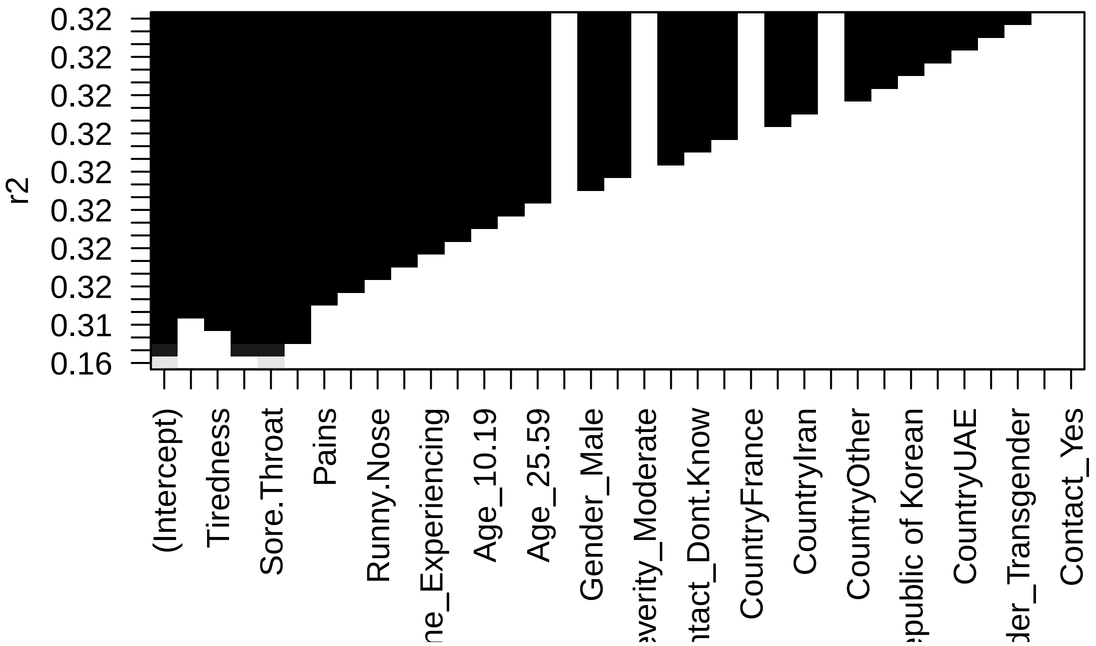
```
plot(covid.summary$bic,xlab="No. of variables",ylab="BIC",type="l")
minBIC = which.min(covid.summary$bic)
points(minBIC,covid.summary$bic[minBIC],col="red",cex=2,pch=20)
```



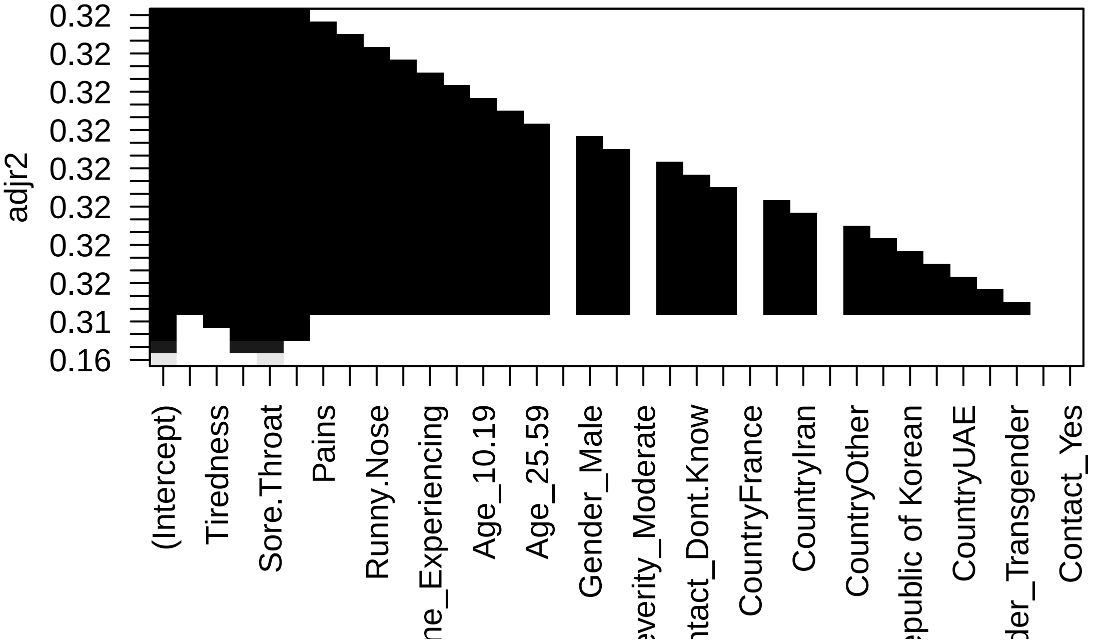
```
paste("Max Rsq: ",maxRsq," Min Cp: ",minCp," Min BIC: ",minBIC)
```

```
## [1] "Max Rsq:  5  Min Cp:  5  Min BIC:  5"
```

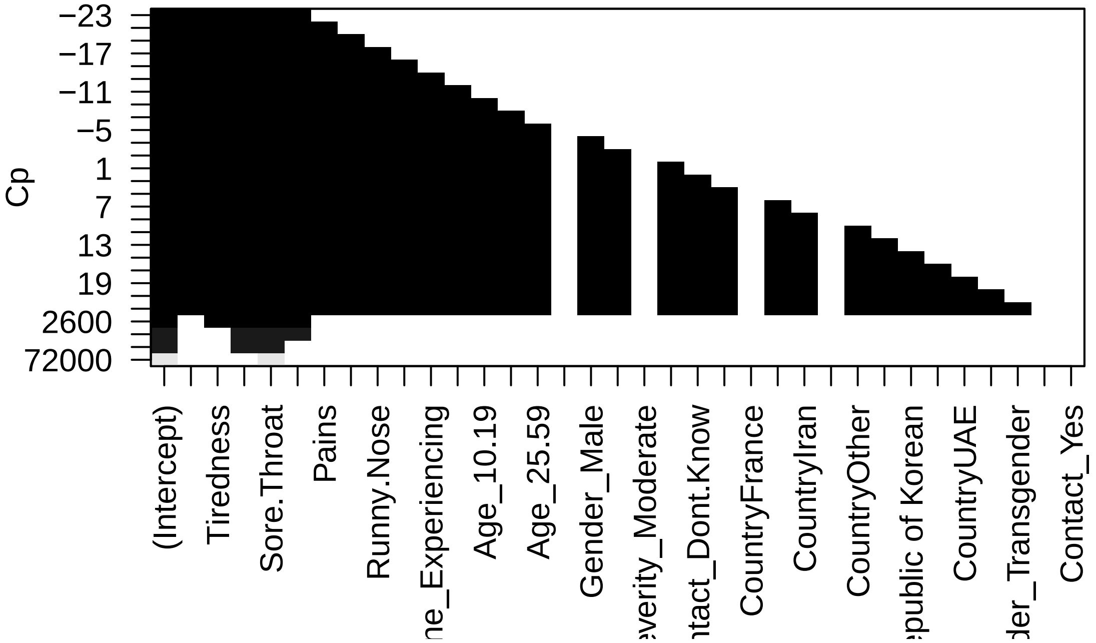
```
# par(mfrow=c(2,2))
plot(regfit.full,scale="r2")
```



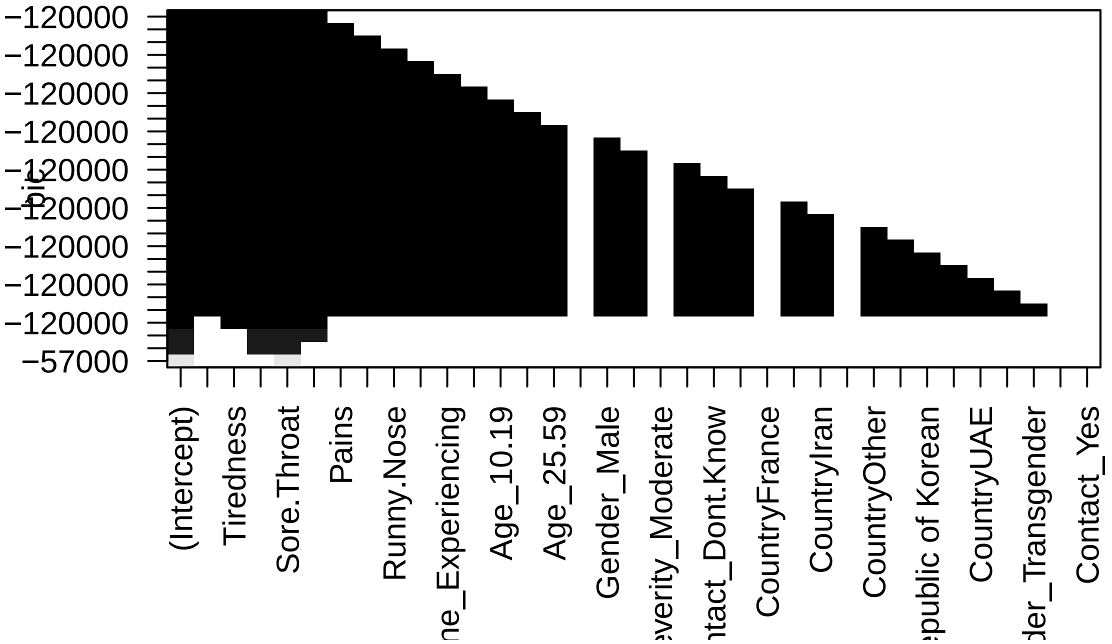
```
plot(regfit.full,scale="adjr2")
```



```r
plot(regfit.full,scale="Cp")
```



```r
plot(regfit.full,scale="bic")
```

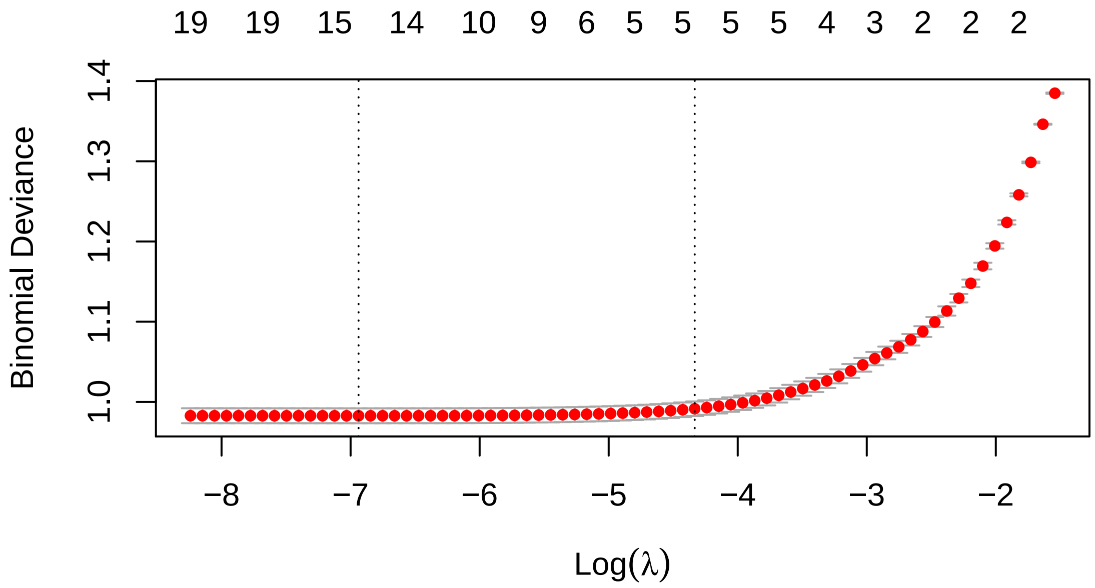


```
#5 variable model has least bic
coef(regfit.full,5)
```

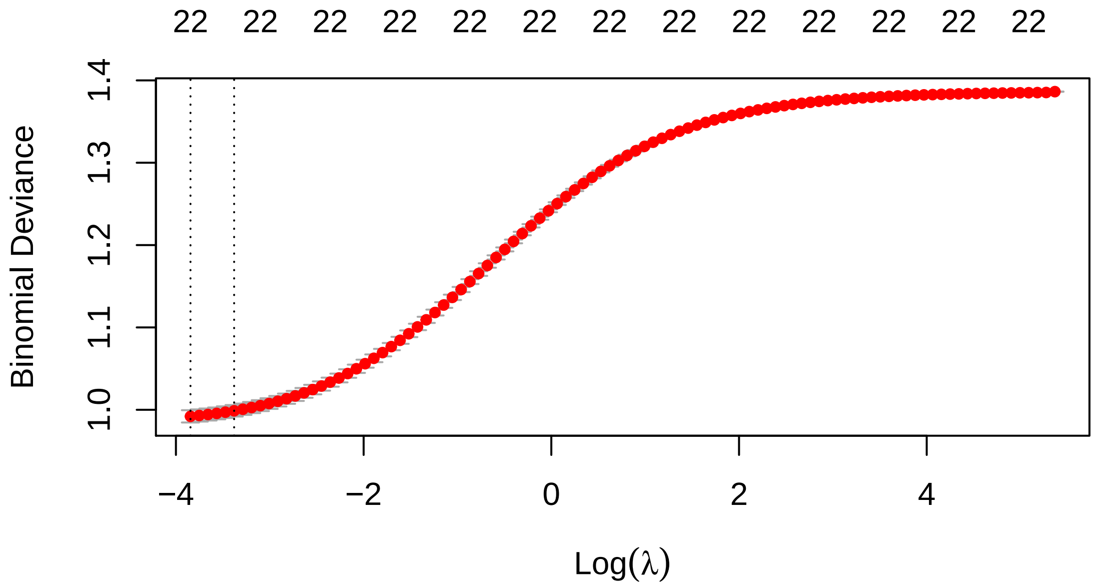
```
##   (Intercept)        Fever     Tiredness     Dry.Cough   Sore.Throat None_Sympton
##    0.27272727  -0.09090909  -0.09090909    0.36363636    0.36363636  -0.27272727
```

```
######################################################
#RIDGE AND LASSO

#taking a smaller set
covid_19_data = covid_19_data[0:10000,]

# Split the data into training and test set
set.seed(123)
training.samples <- covid_19_data$Difficulty.in.Breathing %>%
  createDataPartition(p = 0.8, list = FALSE)
train.data  <- covid_19_data[training.samples, ]
test.data <- covid_19_data[-training.samples, ]


# levels of each var[Taking only predictors with level>=2]
# map(map(train.data,as.factor),levels)
train.data <- train.data[map(map(map(train.data,as.factor),levels),length)>1]
test.data <- test.data[map(map(map(test.data,as.factor),levels),length)>1]
```

```r
x <- model.matrix(Difficulty.in.Breathing ~. , train.data)
y <- train.data$Difficulty.in.Breathing

# Find the best lambda using cross-validation
set.seed(123)
cv.lasso <- cv.glmnet(x, y, alpha = 1, family = "binomial")
cv.ridge <- cv.glmnet(x, y, alpha = 0, family = "binomial")
# par(mfrow=c(2,2))
plot(cv.lasso)
```



```r
plot(cv.ridge)
```



```
#lambda.1se ->lies within 1 standard variation of lambda.min, it gives good accuracy with minimum predi

# Fit the final model on the training data
model.ridge <- glmnet(x,y,alpha=0,family="binomial",lambda = cv.ridge$lambda.1se)

model.lasso <- glmnet(x, y, alpha = 1, family = "binomial",lambda = cv.lasso$lambda.1se)

# Display regression coefficients
coef(model.lasso)
```

```
## 24 x 1 sparse Matrix of class "dgCMatrix"
##                             s0
## (Intercept)         -1.3205516
## (Intercept)          .
## Fever               -0.3523249
## Tiredness           -0.2616559
## Dry.Cough            1.8010549
## Sore.Throat          1.9403445
## None_Sympton        -1.4912053
## Pains                .
## Nasal.Congestion     .
## Runny.Nose           .
## Diarrhea             .
## None_Experiencing    .
## Age_0.9              .
## Age_10.19            .
## Gender_Female       .
## Gender_Male         .
## Gender_Transgender  .
## Severity_Mild       .
## Severity_Moderate   .
## Severity_None       .
## Severity_Severe     .
## Contact_Dont.Know   .
## Contact_No          .
## Contact_Yes         .
```

```r
coef(model.ridge)
```

```
## 24 x 1 sparse Matrix of class "dgCMatrix"
##                              s0
## (Intercept)        -1.078848885
## (Intercept)         .
## Fever              -0.430014755
## Tiredness          -0.302545126
## Dry.Cough           1.605567062
## Sore.Throat         1.680570400
## None_Sympton       -1.747873842
## Pains              -0.009607277
## Nasal.Congestion   -0.001743552
## Runny.Nose         -0.029910944
## Diarrhea            0.019008410
## None_Experiencing   0.049092199
## Age_0.9             0.038423610
## Age_10.19          -0.038342285
## Gender_Female       0.001603919
## Gender_Male         0.008577520
## Gender_Transgender -0.014750825
## Severity_Mild       0.041335234
## Severity_Moderate  -0.004483796
## Severity_None      -0.016224047
## Severity_Severe    -0.019938856
## Contact_Dont.Know   0.002765464
## Contact_No          0.003537363
## Contact_Yes        -0.006275867
```

```r
#FEVER, TIREDNESS DRYCOUGH SORETHROAT NONESYMPTON are most significant according to lasso.
#Whereas RIDGE includes all the parameters
# Make predictions on the test data LASSO
x.test <- model.matrix(Difficulty.in.Breathing~.,test.data)
probabilities <- model.lasso %>% predict(newx = x.test)
predicted.classes <- ifelse(probabilities > 0.5, 1, 0)

# Model accuracy LASSO
observed.classes <- test.data$Difficulty.in.Breathing
mean(predicted.classes == observed.classes)
```

```
## [1] 0.702
```

```
#MSE
mean((predicted.classes- observed.classes)^2)
```

```
## [1] 0.298
```

```
# Make predictions on the test data RIDGE
x.test <- model.matrix(Difficulty.in.Breathing~.,test.data)
probabilities <- model.ridge %>% predict(newx = x.test)
predicted.classes <- ifelse(probabilities > 0.5, 1, 0)

# Model accuracy RIDGE
observed.classes <- test.data$Difficulty.in.Breathing
mean(predicted.classes == observed.classes)
```

```
## [1] 0.695
```

```
#MSE
mean((predicted.classes- observed.classes)^2)
```

```
## [1] 0.305
```

```
###FULL REGRESSION MODEL
# Fit the model
full.model <- glm(Difficulty.in.Breathing ~. , data = train.data, family = "binomial")

# Make predictions
probabilities <- full.model %>% predict(test.data, type = "response")
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
full_predicted.classes <- ifelse(probabilities > 0.5, 1, 0)

# Model accuracy
full_observed.classes <- test.data$Difficulty.in.Breathing
mean(full_predicted.classes == full_observed.classes)
```

```
## [1] 0.687
```

```
#MSE
mean((full_predicted.classes - observed.classes)^2)
```

```
## [1] 0.313
```

```
###############BOOTSTRAP##########################

covid_19_data = read.csv("C:\\Users\\91828\\Documents\\Rlab\\Covid\\Cleaned_Data.csv", header = TRUE)

covid_19_data = covid_19_data[0:10000,]
library(boot)
```

```
##
## Attaching package: ’boot’

## The following object is masked from ’package:lattice’:
##
##     melanoma
```

```
library(glmnet)
library(tidyverse)
library(caret)
# Load the data and remove NAs
covid_19_data <- na.omit(covid_19_data)

# Split the data into training and test set
set.seed(123)
training.samples <- covid_19_data$Difficulty.in.Breathing %>%
  createDataPartition(p = 0.8, list = FALSE)
train.data  <- covid_19_data[training.samples, ]
test.data <- covid_19_data[-training.samples, ]

# levels of each var[Taking only predictors with level>=2]
# map(map(train.data,as.factor),levels)
train.data <- train.data[map(map(map(train.data,as.factor),levels),length)>1]
test.data <- test.data[map(map(map(test.data,as.factor),levels),length)>1]

alpha.fn=function(data,index){
  X=data$X[index]
  Y=data$Y[index]
  return((var(Y)-cov(X,Y))/(var(X)+var(Y)-2*cov(X,Y)))
}
# boot(data=train.data,statistic = alpha.fn,R=1000)
boot.fn = function(data,index){
  return(coef(lm(Severity_Severe~.,data=data,subset=index)))
}

boot.fn(data=train.data,1:100)
```

```
##             (Intercept)                   Fever               Tiredness
##            1.000000e+00                      NA                      NA
##               Dry.Cough Difficulty.in.Breathing             Sore.Throat
##                      NA                      NA                      NA
##            None_Sympton                   Pains        Nasal.Congestion
##                      NA           -2.665145e-16           -3.460276e-16
##              Runny.Nose                Diarrhea       None_Experiencing
##           -8.127886e-17           -2.013254e-16           -1.381457e-15
##                 Age_0.9               Age_10.19           Gender_Female
##                      NA                      NA                      NA
##             Gender_Male      Gender_Transgender           Severity_Mild
##                      NA                      NA           -1.000000e+00
##       Severity_Moderate           Severity_None       Contact_Dont.Know
##           -1.000000e+00           -1.000000e+00           -5.965617e-17
##              Contact_No             Contact_Yes
##           -1.747442e-16                      NA
```